DOCUMENT RESUME

ED 265 166 TM 850 501

AUTHOR Yap, Kueh Chin; Capie, William
TITLE The Influence of Same Day or Separate Day Observations on the Reliability of Assessment Data.
PUB DATE Apr 85
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; *Classroom Observation Techniques; Data Collection; Elementary Secondary Education; Field Tests; *Generalizability Theory; Interrater Reliability; Teacher Behavior; *Teacher Evaluation; *Test Reliability; True Scores
IDENTIFIERS *Teacher Performance Assessment Instruments

ABSTRACT

The purpose of this study was to compare the relative magnitude of the variance components and generalizability coefficients derived from the Teacher Performance Assessment Instruments (TPAI) data using two different methods of data collection: (1) occasions when observers were in the classroom for simultaneous observation and (2) occasions when the observers made independent observations. The two data sets used in the analysis were collected as part of the field testing of the revised TPAI, and they consisted of two different sets of observations of 23 teachers. One set consisted of observations by two different observers present in the teachers' classrooms on the same day; the other set consisted of observations of these same teachers by two different observers on different days. Observations from each of these data sets were scored using the criteria anticipated when the instruments are used in certification, including the use of essential descriptors. Satisfactory values for the generalizability coefficient were obtained when two observers visited the classroom on separate days, but a single day of observation did not provide sufficiently reliable data to differentiate teachers. (LMO)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

The Influence of Same Day or Separate Day Observations on the Reliability of Assessment Data

ED265166

Kueh Chin Yap
University of Georgia

William Capie
University of Georgia

April 1985

Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois

TM 850 501

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Kueh Chin Yap

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Influence of Same Day or Separate Day Observations on the Reliability of Assessment Data

The generalizability of measures of teaching has been a concern for some time. The issue has been examined in a number of different ways by various investigators who were concerned with the stability of behavior, number of observers, the amount of observation time and other similar problems. (See for example, Shavelson and Dempsey-Atwood, (1976); Rowley (1976 and 1978); and Capie, 1981). Stodolsky (1984), in citing some of these earlier works, questioned whether teaching behaviors are sufficiently stable to be used in teacher evaluation programs such as those in Georgia and South Carolina. This question would have a great deal of common sense credibility to teachers who use more management moves on a Friday afternoon than they do on a typical Tuesday morning. However, the question must be explored and answered in light of the way in which the evaluation system deals with the varying contexts and then converts the data into scores or other evaluation decisions.

The Teacher Performance Assessment Instruments (TPAI) put the first steps of this process into the hands of experienced teachers and administrators who are trained as observers. These observers do not attempt to count or to otherwise record particular behaviors. Instead, they are charged with making some one hundred and forty decisions like this: "In the context of the lesson (i.e. type of activity, level of learners, etc.), did the teacher use sufficient non-verbal techniques to redirect off-task learners?" Thus, with a well-behaved group a small number of behaviors would be considered sufficient, while more may be

required on a Friday afternoon. These 140 decisions that an observer makes are then aggregated into nine competency decisions using scoring rules laid out in the instrument. With the context variations and potential instability addressed in this way, concern for the generalizability of behaviors is replaced with the more relevant problem of the generalizability of the scores or the generalizability of the evaluation decisions.

The process of assessing the generalizability of the scores is the same as assessing the generalizability of the behaviors. First, there should be concern for the stability of the results from one occasion to another. Thus, if a teacher does a superior job in two very different contexts, the evaluation decision should be the same even though the frequencies of particular behaviors will be quite variable. Secondly, there should be generalizability across observers or interobserver agreement. A third type of generalizability which involves the items, or constituent behaviors that contribute to a decision, is the internal consistency of the measures. There might also be concern for intraobserver agreement, but the concept is difficult to assess except artifically when observers can review the same videotaped lesson.

Generalizability analyses of the TPAI have taken into account each of these three sources of variation since there were facets representing indicators, types of observers, and assessment occasions (fall and spring). The identity of the individual observers, variation in school levels, and a host of other context variables all were masked or confounded in the residual term where they surpress the reliability.

The TPAI certification data that have been used in generalizability analyses have actually contained two occasion effects. The first is easily identified and assessed since it is associated with fall and spring differences. The second occasion effect is more difficult to assess since it is confounded with observer effects. Since observers go into classrooms on different days and see different lessons, it is not possible to determine if variation in the scores is a result of low interobserver agreement or a result of variations in the teachers' performance from lesson to lesson. Past analyses of the TPAI have been based on the assumption that all of this ambiguous variation is observer error. While having variance associated with an unstable set of scores may or may not be any more or less desirable than having observer error variance, there is some interest in knowing how much of the apparent observer variance may represent instability of behavior.

PURPOSE

The purpose of the study was to compare the relative magnitude of the variance components and generalizability coefficients derived from TPAI data using two different methods of data collection: 1) occasions when observers were in the classroom for simultaneous observation and 2) occasions when the observers made independent observations.

PROCEDURES

The two data sets used in the analysis were collected as part of the field testing of the revised TPAI. These data sets consisted of two different sets of observations of twenty-three teachers. One set consisted of observations by two different observers present in the teachers' classrooms on the same day.

The other data set consisted of observations of these same teachers by two different observers on different days.

The teachers were volunteers who agreed to prepare a lesson plan portfolio and have observers come into their classes. Observations from each of the data sets were scored using the criteria anticipated when the instruments are used in certification, including the use of essential descriptors. Raw and transformed scores were computed utilizing the methods outlined in the Cronin and Capie paper (1985).

ANALYSIS

For each of the data sets, two types of analyses were conducted.

Pass-Rate

For each data set, the twenty-three teachers were scored on all levels of the assessment instruments. The portion of indicators scored acceptably by the observers constituted the "score" for both data sets.

Dependability

Generalizability theory was used to plan the analysis of the field test data. Four facets were identified as important sources of variation in the performance data obtained: teachers; observers; observer-types; and performance indicators. The four facet design with observers nested within observer-type is identical to a three facet fully crossed design with teachers, observer-types, and performance indicators as the sources of variation. As a consequence, the simpler three facet model was used in all analyses.

For each analysis, teachers were treated as facets of differentiation while observers and performance indicators within competencies were treated as facets of generalization. All facets were regarded as random in the analysis design.

A dependability coefficient (ϕ) was calculated to assess the dependability of the data for making judgements about teacher performance relative to a standard (λ). According to Brennan (1978):

$$\hat{\phi}(\lambda) = \frac{\hat{\sigma}^2(T) + (\bar{\chi} - \lambda)^2 - \hat{\sigma}^2_{\bar{\chi}}}{\hat{\sigma}^2(T) + (\bar{\chi} \quad \lambda)^2 - \hat{\sigma}^2_{\bar{\chi}} \hat{\sigma}^2_{\Delta}}$$

$\hat{\sigma}(T)$ = the variance attributable to the facet of differentiation:

$\hat{\sigma}^2_{\bar{\chi}}$ = the variance components of the observed mean scores;

$\bar{\chi}$ = the mean score;

λ = the cutoff score;

$\hat{\sigma}^2_{\Delta}$ = variance component due to error

RESULTS

The mean number of indicators mastered by the 23 teachers is shown in Table 1. When observations were made on the same day the mean was 28 or 80 percent of the indicators at-or-above the minimum standard. When observers were done on different days the mean number of indicators scored acceptably was 26.8 or 76.5 percent.

Insert Table 1 about here

The variance components associated with the analyses are included in Table 2.

Insert Table 2 about here

The largest value of rho squared, .599, was associated with acceptable/non-acceptable data on different days.

DISCUSSION

Consideration of these results must be made in light of the context in which the study was done. The TPAI used was a preliminary field test edition which had had limited use prior to the study. Furthermore, the observers had had little or no training in its use prior to interpretation. The RAC team members had participated in reviews of earlier drafts and had a brief orientation meeting. The school site observers may have had an orientation meeting. Such arrangements were tolerable in a field-testing situation since the assessments were followed by extensive debriefing to identify problems which required attention in instrument revision and/or in training.

An additional factor that limits the direct applicability of these results is that the analyses have been conducted at the instrument level rather than at the competency level where decisions are made. This does not limit the heurstic value of the findings, but the findings may differ from one competency to another.

The purpose of the study was to compare the reliability coefficients and the variance components associated with various factors in teacher evaluations when two observers visited the classroom together or independently. Seven sources of variation are each discussed for the analyses of raw and transformed data.

Raw Data

Variance due to teachers is desirable, of course, and necessary to differentiate teachers. In these two situations the variance due to teachers when observed on different days was about three times that of teachers observed on the same day.

Variance due to type of observer is not desirable since there should be no differences in the decisions made by trained observers. In this case the variances were near zero for both situations.

Variance due to indicators is to be expected to the extent that teachers, as a group, find certain indicators easier or more difficult than others. In this case the variance due to indicators was quite substantial and almost equal for both cases.

The teacher by observer interaction is one which is particularly pertinent to the study. It is an indication of the extent to which certain teachers may have been seen differently by one set of observers or the other. In these analyses, where

individual RAC staff members or school site persons are confounded with observer type, it is actually an indication of the extent to which one observer sees a teacher differently than the other. In cases when the observers are seeing the same lesson, this variance component is an indication of the extent to which the observers' overall impressions of the teacher differ. In cases when the observers see different lessons, this variance component is a combination of this factor plus differences in the two lessons observed. With the raw data the two variance components were .211 on different days and .071 on same day observations

The teacher by indicator interaction is an indication of the extent to which particular teachers experience relative ease or difficulty with certain indicators. There was a relatively large difference in the magnitude of these coefficients associated with the two methods of observation.

The observer by indicator interaction is an indication of how much difference there is in the way that the two observer types score particular indicators. A value of near zero would suggest that there is little difference in the way the school-site observers and off-site observers scored each indicator. The values associated with the raw scores suggest that there are few differences in the way the two groups see particular indicators.

The final variance component represents teacher by observer by indicator interactions. Sometimes designated as the residual, it is an indication of the degree to which individual indicators are scored differently by the two observers of individual teachers. Thus, when two observers see the same lesson, it is an indication of interobserver agreement or disagreement on

particular indicators. When the observers see different lessons, this variance component is a combination of these types of observer differences and variations in the teacher's performance. With the raw scores there was not too much difference in these two values which suggests that there may have been reasonable agreement when observers viewed the same lessons.

Transformed Scores

Variance due to teachers in transformed scores was quite different when the two observers observed on two days rather than on the same day. The magnitude of the variance component associated with two observation days is similar to that in previous studies, however. Needless to say, if very little of the variance is associated with teachers, there will be low reliability.

There was no variance associated with observer effect for either type of observation schedule. This is a desirable finding since it suggests no systematic leniency of one observer type.

Variance due to indicators was approximately equal with the two situations. This area might be somewhat inflated by the pilot context where teachers did not feel compelled to give maximum effort to difficult aspects of planning.

The teacher by observer interaction, which is particularly important, showed differences when two observation days were compared with one. The two day variance component which includes variance due to varying teacher behavior and variance due to different observers, was substantially larger than the single day variance component. This effect, which had been expected, is highly significant because it may carry the implication that the

teacher by observer variation had not been as laden with observer error as prior analyses had assumed.

The magnitude of the teacher by indicator interaction variance component was similar in the two observation models.

The observer by indicator interaction variance component was zero in both cases as it should be. The residual or three way interaction variance component was approximately equal in both situations.

The Generalizability Coefficients

Although the variance components representing error, notably teacher by observer and teacher by observer by indicator interactions, did behave as expected, some of the possible gains in the magnitude of the reliability coefficient were offset by small variation among teachers observed on a single day. Other than the obvious fact that once may not be enough, reasons for the low variance on a single day are not apparent. Capie, Ellett and Cronin (1985) have reported somewhat different results after comparing the generalizability of scores from multiple observers on a single day with those of multiple observers each on separate days. However, the two sets of teachers in their study were not identical. It may have been that their results would have been different if they had observed both teacher groups on both observation schedules.

A single study could be designed to address the issues raised in this and the Capie, Ellett and Cronin study. However, it would be quite costly, requiring multiple observers observing multiple teachers on multiple occasions. Furthermore, it may be desirable to contemplate having several on-site and off-site observers to

allow for exploring various combinations of observers. As Capie, Cronin and Yap (1985) showed, (in the previous paper), the nature of the observers does matter. Computer programs can add another theoretical observer. But, that person must be either on-site or off-site-not two-thirds of one and one-third of the other. A further complication in designing the "ideal" study is the problem with the stability of the variance components. Tobin and Capie (1981) have claimed that more than fifty teachers may be required with data such as these where the number of observers and indicators is fixed. Given the practical difficulties in designing this ideal study, it behooves us to speculate slightly about the interpretation of these results.

Two variance components in this design are particularly important in reflecting possible instability. Interestingly, these same two variance components, teacher by observer and teacher by observer by indicator, constitute the error term when computing the generalizability coefficient.

With the variance components for the transformed data gathered on multiple days, reducing the teacher by observer variance component by half reduces the error term by forty-four percent. Meanwhile, reducing the teacher by observer by indicator interaction reduces the error term by only 7 percent. Clearly, any reduction in the size of the error term will bring an increase in the magnitude of ρ^2.

Of course, as the number of indicators is decreased the reductions in the size of the residual term becomes more important and reductions in the teacher by observer term have proportionally less impact.

Such speculation leads to the obvious question of how much of the apparent observer variance really is instability. In this study the difference reduction was by a factor of three(.016 to .005). In the Capie, Ellett, Cronin (1985) data it was .022 for field test data and .005 for training data.

The simulation that was presented assumed a conservative 50 percent reduction in variance related to stability. Clearly, the portion of the variance that can be associated with observers who are scoring the same lesson is error, whether it be due to personal biases or even instrument inadequacies. With an improved instrument, trained observers, etc. this kind of error should be reduced. In future field testing, competency by competency analyses may provide indications of where training or instrumentation has been inadequate. Past experience has shown that teacher variance will also diminish with real assessment pressure, however (Capie & Ellett, 1982).

The significance of the variance that "disappears" when observation is done on only one day is an interesting issue Stodolsky (1984) has suggested that with behavioral measures this variance is a kind of error which may be a limiting factor in the generalizability of teacher evaluations. Yet, even with a field test version of the TPAI, ρ^2 values of .7-.9 were reported. Of course, the TPAI are aggregated evaluation decisions rather than behavioral frequencies. Nevertheless, the sample did not consist of superior teachers. Instead, there were teachers who were variable in the adequacy of their performance. Including this stability factor, which may be quite sizeable, in the error term

is a conservative approach. However, there isn't sufficient justification right now for choosing a different approach.

The results show that some care must be taken in structuring the data collection system since the generalizability coefficients were satisfactory when there were multiple observers observing on separate days but not for multiple observers on the same day. Care must be taken also, in the design of the reliability analysis lest these effects be overlooked.

CONCLUSION

Satisfactory values for the generalizability coefficient, ρ^2, were obtained when two observers visited the classroom on separate days. A single day of observation did not provide sufficiently reliable data to differentiate teachers. When observations were made on the same day relatively small variance components were observed for factors logically related to day-to-day variation. The difference was attributed to day-to-day variation in the quality of particular teaching behaviors. Consequently, the analyses which included this teacher instability as observer error were considered to be conservative.

References

Brennan, Robert L. (1978). Extensions of generalizability theory to domain referenced testing, American College Testing Program, Iowa City.

Capie, W. (1981). Sampling considerations in classroom research. A paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Capie, W., & Ellett, C. (1982). The effects of assessment demand characteristics on the dependability of teacher performance measures. A paper presented at the annual meeting of the American Educational Research Association, New York.

Capie, W., Cronin, L., & Yap, K. C. (1985). The influence of observer type on assessment decisions and their reliability. A paper presented at the annual meeting of the American Educational Research Association, Chicago.

Capie, W., Ellett, C., & Cronin, L. (1985). Assessing meritorious teacher performance: Reliability, decision-making, and standards setting procedures. A paper presented at the annual meeting of the American Educational Research Association, Chicago.

Cronin, L. and Capie, W. (1985). The influence of scoring procedures on assessment decisions and their reliability. A paper presented at the annual meeting of the American Educational Research Association, Chicago.

Rowley, G. L. (1976). The reliability of observational measures. American Educational Research Journal, 13(1), 57-59.

Rowley, G. (1978). The relationship of reliability in classroom research to the amount of observation: An extension of the Spearman-Brown formula. Journal of Educational Measurement, 15(3), 165-180.

Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. Review of Educational Research, 46(4), 553-611.

Stodolsky, S. (1984). Teacher evaluation: The limits of looking. Educational Researcher, 13, 11-18.

Table 1

Mean Number of Indicators Mastered
(I=35)

Source	Mean	Percent
Two Observers on Same Day	28	80
Two Observers on Different Days	26.8	76.5

Table 2

Variance Components
(T=23, O=2, I=35)

Source	Acceptable/Not Acceptable 1/0 Scoring		Raw Scores	
	Same Day	Different Days	Same Day	Different Days
Teacher(T)	.003	.015	.039	.121
Observer(O)	.000	.001	.004	.000
Indicator(I)	.035	.032	.443	.417
TO	.005	.016	.071	.211
TI	.027	.025	.348	.160
OI	.000	.000	.000	.012
TRI	.090	.100	.683	.845